

Big data and Hadoop.

In this video, we're going to cover big data, and what it is, and why we should use it.

Then, we'll introduce Apache Hadoop and what it means to big data.

These are the three primary drivers for big data.

Increasing data volumes, the annual growth of www information volume is 59% and it continues to rise.

This explosion of new data,

is driven by a full range of traditional and non-traditional sources like sensors, devices, bots and crawlers.

Increasing data and analysis complexity.

The real growth of data mentioned earlier, is coming from unstructured data.

And the myth that 80% of unstructured data has no value has been debunked by examples like the success of search engine providers and

e-retailers who unlock the value of click stream data.

The requirement to store, analyze and mine structured and unstructured data together is becoming the new norm.

Changing economics and emerging technologies.

Cloud computing and commodity hardware have radically reduced the acquisition cost of computational and storage capacity, and

is fundamentally changing the economics of data processing.

Commodity hardware is being complemented by new distributed parallel processing frameworks like Hadoop which when combined with the rich ecosystem of tools, provides a platform for tackling massive data processing tasks.

Each of these drivers is especially relevant considering that we're living in a world where data is key.

Where information leads to insights and forms all of our business decisions.

There are a new set of questions that organizations are asking to drive their business.

Answer these new types of questions and you'll be able to more effectively differentiate and derive new value for the business.

Whether it is in the form of revenue growth or cost savings or creating entirely new business models.

Today new types of questions are being asked.

Some of these questions include questions on social and web analytics.

What is my brand and product sentiment?

How effective is my online campaign?

And who am I reaching?

How can I optimize or target the correct audience?

Then there are questions that require connecting to live data feeds.

A large shipping company uses live weather feeds and traffic patterns to fine tune its ship and truck routes, leading to improved delivery times and costs savings.

Retailers analyze sales, pricing, and economic, demographic, and live weather data to tailor product selections at particular stores, and to determine the timing of price markdowns.

Questions that require advanced analytics.

Financial firms using machine learning to build better fraud detection algorithms that go beyond the simple business rules involving charge frequency and location to also include an individual's customized buying patterns.

Ultimately leading to a better customer experience.

What do we mean by big data?

Big data represents the trends, technologies, and potential for businesses to obtain valuable insights from large, meaning petabytes or even exabytes, amounts of structured, unstructured, and fast moving data.

ERP, SCM, CRM, and transactional web applications are classic examples of systems processing transactions.

Highly structured data in these systems is typically stored in SQL databases.

Web 2.0 is about how people and things interact with each other, or with your business.

Web logs, user clickstreams, social interactions and feeds and user generated content or classic places to find interaction data.

In the area of ambient data and Internet of Things, Merrymaker has predicted 10 billion connected devices by 2015.

Sensors for heat and motion, pressure, and RFID and GPS chips within such things as mobile devices, ATM machines, and even aircraft engines provide just some examples of things that output ambient signal.

There are multiple types of data, personal, organizational, public and private.

So we should not minimize our thinking to just data that flows through an organization.

Organization data you may have could benefit from being blended with external data.

Moreover, the government has the Open Data Initiative.

Which means that more and more data is becoming publicly available.

YouTube videos, Facebook posts, credit card transactions, store inventory, your last grocery purchase.

Trillions of pieces of information are being collected, stored and analyzed almost daily with increasing speed.

Big data addresses one of the most critical issues facing businesses today.

How to gain value from the growing realms of complex data.

Big data expands the capabilities of traditional BI to derive meaningful new insights from new types of data.

Characterized by high volume, velocity and variety.

Faster, better and

more comprehensively at a lower cost than previously possible.

It represents the trends, technologies and potential for

businesses to obtain valuable insights from large,

again, petabytes or even exabytes amounts of structured, unstructured and fast-moving data.

Big data technologies along with the BI capabilities can promote visibility into your data and

systems for applications to increase the richness and variety of the big-data experience.

Along with the analysis tools,

that can help reveal actionable insights into opportunities, execution, performance and impact on the business.

Some of the example scenarios are ITM for structure optimization.

Develop new innovative IT services to enable internal and external business clients and

streamline IT operations and modernize technology platforms to exploit today's opportunities.

Social network analysis.

Analysis of social data and

sentiment analysis to drive customer insight.

Similarly other scenarios are fraud detection,

churn analysis and many more.

This is an opportunity to gain competitive advantage by moving first and fast in your industry.

Everyone has heard of Hadoop but what is it and why do I need it?

Apache Hadoop is an open source solution framework

that supports data-intensive distributed applications on large clusters of commodity hardware.

Any organization looking at their current data

infrastructure and seeing a significant shift in the volume, variety and velocity of data is experiencing big data.

Data volume is growing by ten times every five years.

In other words, there may be oceans of data out there, but making it into something that you can use

particularly in real time is another matter entirely.
The variety of data is also challenging for additional data storage technologies.
Up to 85% of new data is unstructured such as Facebook posts and tweets, videos, photos, search engine queries, etc.
Finally, the velocity of data creates new challenges as organizations now nearly instantaneously capture vast amounts of data from websites, machine-generated data like sensors on smart grids, point-of-sale devices and other sources.
We see the real breakthrough insights coming through when we take what is the traditional business intelligence and add more capability such as machine learning, predictive analysis, statistical analysis.
Large scale graph processing, pattern mining, trend analysis and economic modelling.
All of which today are a reality in Hadoop.
But there is some misunderstanding about Hadoop. Let's tackle those misconceptions by outlining what Hadoop is not.
Hadoop is a compliment, not a replacement, as seen in the view of an enterprise data warehouse. It acts as yet another cog in the data analytics that must play well with others.
Hadoop is not a place to learn how to code.
There are currently a lot of assumptions in the Hadoop APIs and documentation.
Assumptions that you know the basics of Java programming and the common error messages you get when things don't work.
But if you do not know about class paths, how to compile and debug Java code, step back from Hadoop and learn a bit more about Java before proceeding.
Finally, HDFS is optimized to access batches of data sets quicker meaning high throughput rather than particular records in that data set, an example of low latency.
Applications that require a little latency access to data in the tens of milliseconds range, will not work well with Hadoop.
Here are some industry scenarios of Hadoop.
Financial services, modelling true risk and threat analysis, fraud detection and trade surveillance, credit scoring and analysis, retail, point of sales transaction analysis, customer churn analysis and sentiment analysis.
Telecommunications, customer churn prevention, network performance optimization, call detail record analysis, and network failure prediction.

Healthcare, genomics research, cancer research, health pandemics early detection, and air quality monitoring.

What is Hadoop?

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Rather than rely on hardware to deliver high availability, the library itself is designed to detect and handle failures at the application layer.

This means it delivers a highly available service on top of a cluster of computers, each of which may be prone to failures.

The base Apache Hadoop framework is composed of the following modules.

Hadoop Common, which contains libraries and utilities needed by other Hadoop modules.

Then there's Hadoop Distributed File System or HDFS which is a distributed file system that stores data on commodity machines providing very high aggregate bandwidth across the cluster.

Hadoop MapReduce is an implementation of the MapReduce programming model for large scale data processing.

Then there's Hadoop YARN which is a resource management platform that is responsible for managing computing resources and clusters and using them for scheduling of user's applications.

In addition to the core,

Hadoop has a rich ecosystem that supports SQL/NoSQL, streaming and real-time and interactive applications.